Рубаков Сергей Валерьевич
начальник отдела разработки
программного обеспечения
компании Corpus Technologies

СОВРЕМЕННЫЕ МЕТОДЫ АНАЛИЗА ДАННЫХ

Введение

Любые методы обработки данных так или иначе используются для структурирования и анализа существующей информации. Задач по анализу информации много, однако в этой статье рассмотрены методы, которые эффективно работают для решения задач по структурированию данных с большим количеством разнородных параметров.

Например, для более эффективного продвижения товаров массового потребления на рынок имеет смысл сегментировать потребителей на группы по определенным параметрам: пол, возраст, семейное положение, доход семьи и так далее. Для этого существует набор математических методов, которые позволяют установить закономерности в данных – в случае с анализом потребителей такой закономерностью будут характерные группы потребителей. Часто случается, что формальные методы анализа позволяют получить неожиданные новые знания – например, при исследовании клиентов одной из гостиниц выяснилось, что все клиенты-пенсионеры, проживающие в гостинице, имеют доход свыше 800 долларов [1]. Прежде чем мы остановимся на методах анализа данных более подробно, рассмотрим то, какие данные могут быть использованы для анализа и какое количество данных необходимо.

1. Типы данных

Данные, которые могут быть использованы для анализа, бывают четырех типов [2]:

1. Численные данные (стоимость товара; 100 рублей).
2. Интервальные данные (доля рынка компании; 5 %).
3. Ранговые данные (лояльность потребителя; напиток Coca-Cola нравится *больше*, чем Pepsi-Cola).
4. Номинальные данные (профессия потребителя; ученый, военный, врач).

Все данные, которые подходят под один из этих типов, могут быть проанализированы с помощью формальных методов.

Любой набор данных может быть адекватно представлен комбинацией перечисленных типов.

2. Количество данных

Для того чтобы работали большинство методов, желательно иметь более 30 событий (малая выборка) [3].

Этого количества событий обычно достаточно для получения информации, что в данной выборке наблюдается статистический эффект. Однако для разделения на группы необходимо иметь уже гораздо большее число событий – примерно 30, умноженное на число групп.

Например, для более-менее правильного разделения потребителей на 3 группы желательно иметь более 100 респондентов. Несомненно, для разных задач и методов количество событий может быть разным, и какую-то информацию можно извлекать уже из 10 событий, однако здесь действует общее правило статистики: чем данных больше, тем лучше.

3. Методы анализа и обработки данных

3.1. Кластерный анализ

Термин «кластерный анализ» (впервые ввел Tryon в 1939 г. [4]) в действительности включает в себя набор различных алгоритмов классификации. Общий вопрос, задаваемый исследователями во многих областях, состоит в том, как разбить данные на группы с близкими значениями параметров.

Например, при сегментации рынка можно кластеризовать потребителей по двум параметрам — цены и качества. Допустим, компания — производитель автомобилей провела опрос потребителей, в котором задавала два вопроса: «За какую цену Вы готовы купить автомобиль?» и «Оцените качество автомобиля X по 50-балльной шкале» (несколько странный вопрос, однако в качестве иллюстрации он вполне подходит). В результате опроса были получены следующие данные (см. схему; данные в табличной форме не носят информативный характер):

№ участника опроса	Цена, тыс. $	Качество автомобиля X
1	27	19
2	11	46
3	25	15
4	36	27
5	35	25
6	10	43
7	11	44
8	36	24
9	26	14
10	26	14
11	9	45
12	33	23
13	27	16
14	10	47

Схема. Результаты опроса

Если посмотреть на диаграмму (так называемая диаграмма рассеяния) «цена — качество», представленную на рис. 1, то сразу будут видны группы потребителей:

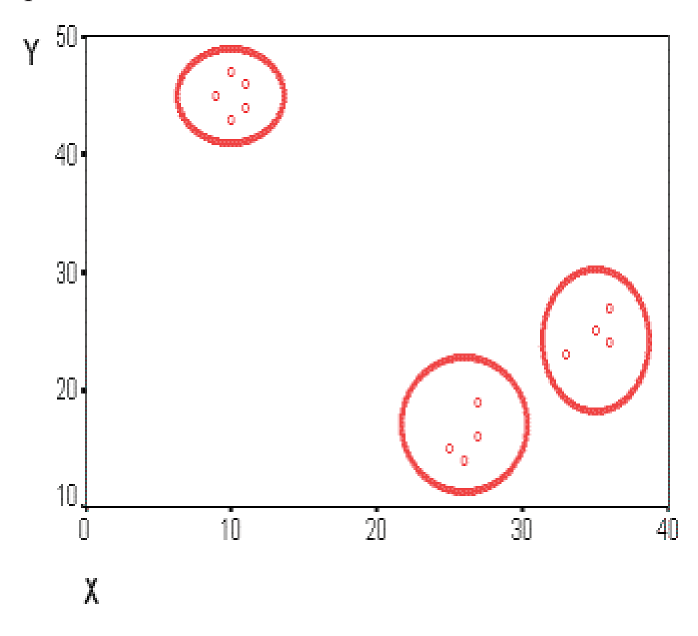


Рис. 1. Соотношение цены — качества

Владея этой информацией, каждой группе потребителей можно предложить именно то, что необходимо именно этой группе, и за счет этого увеличить уровень продаж компании.

Разумеется, в реальной жизни кластеры, различимые глазом, встречаются нечасто, гораздо чаще бывают ситуации, когда все результирующие параметры смешиваются в одну «кучу» (см. рис. 2)

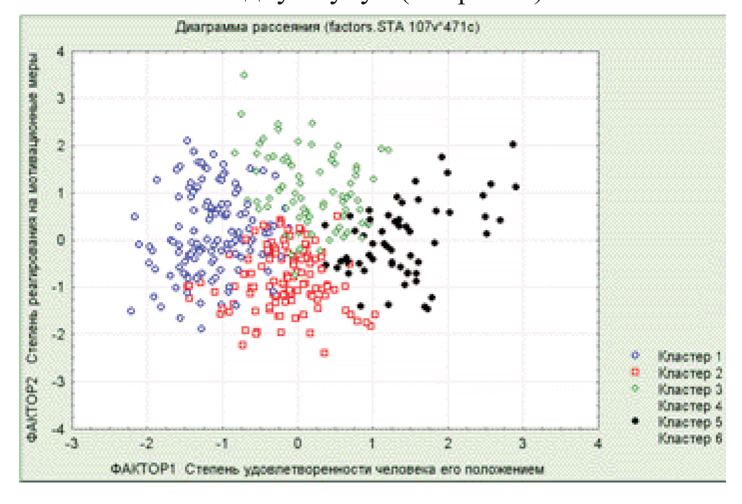


Рис. 2. Диаграмма рассеяния

Особенно часто это встречается, когда анализируемых параметров не два, а несколько десятков (кластерный анализ не ограничивает число анализируемых параметров, поэтому можно рассматривать всю проблему комплексно).

Глазом кластеры выделить не получится, однако с помощью алгоритмов кластерного анализа это сделать можно.

Для проведения кластерного анализа, кроме сбора данных, необходимо определить две вещи: на какое количество кластеров необходимо разделить данные и как определить меру сходства в данных. Например, все предприятия России можно кластеризовать по географическому признаку на 10 кластеров. Тогда мера сходства будет определяться коммуникационной близостью предприятий друг к другу. В более сложных случаях можно применять другие меры сходства, которые подробно описаны в литературе по кластерному анализу [5]. Существует много разных мер сходства, наиболее употребительны из них порядка десяти, но подробно останавливаться на мерах сходства не будем.

3.2. Факторный анализ

В случае наличия большого числа параметров (более 100) имеет смысл сгруппировать параметры и анализировать уже не каждый параметр в отдельности, а группы параметров как единый комплексный параметр (фактор).

В основе факторного анализа лежит идея о том, что за сложными взаимосвязями явно заданных признаков стоит относительно более простая структура, отражающая наиболее существенные черты изучаемого явления, а «внешние» признаки являются функциями скрытых общих факторов, определяющих эту структуру.

Например, для анализа структуры экономического роста России можно проанализировать все макроэкономические параметры, предварительно объединив их в группы. Одним из таких факторов будет являться ВВП [6].

Объединение параметров можно делать вручную, эмпирически, как это сделано с ВВП, а можно с помощью метода факторного анализа. Применение факторного анализа позволяет, во-первых, уменьшать (редуцировать) число рассматриваемых параметров, во-вторых — находить осмысленные группы параметров, каждая из которых будет являться одним самостоятельным параметром.

Спецификой этого метода является то, что при объединении параметров в факторы каждый фактор аккумулирует в себе общие закономерности во всех параметрах, отбрасывая особенности каждого параметра в отдельности.

3.3. Нейронные сети

Начало нейронным сетям как инструменту анализа данных было положено в начале 40-х годов в работе МакКаллока и Питтса [7]. В этой работе предлагалась модель искусственного нейрона.

Предполагалось, что, моделируя нейронную структуру мозга, возможно приблизиться к искусственному интеллекту. К тому времени уже

было известно, что мозг человека состоит из особых биологических клеток – нейронов, и казалось, что построение сетей из нейронов позволит решать сложные задачи, которые ежедневно решает мозг человека.

С тех пор интерес к нейронным сетям периодически то возрастал, то спадал, что обуславливалось новыми разработками в этой области, и сейчас нейронные сети являются одним из достаточно популярных инструментов анализа данных.

Нейронные сети могут быть применены практически к любой области деятельности, что сильно привлекает многих исследователей.

3.3.1. Задачи, которые ставятся перед нейронными сетями

По мнению Anil K. Jain из Мичиганского государственного университета и специалистов Исследовательского центра IBM Jianchang Mao и K. M. Mohiuddin, список задач для нейронных сетей можно классифицировать следующим образом [8].

Классификация образов. К известным приложениям относятся распознавание букв, распознавание речи, классификация сигнала электрокардиограммы, классификация клеток крови, обеспечение деятельности биометрических сканеров и т. п.

Кластеризация / категоризация. Кластеризация применяется для извлечения знаний, сжатия данных и исследования свойств данных.

Аппроксимация функций. Типичным примером является шумоподавление при приеме сигнала различной природы, вне зависимости от передаваемой информации.

Предсказание / прогноз. В качестве примера можно привести предсказание цен на фондовой бирже и прогноз погоды.

Оптимизация. Назначение штата работников по ряду умений и факторов являются классическими примерами задач оптимизации.

Память, адресуемая по содержанию (ассоциативная память). Ассоциативная память доступна по указанию заданного содержания. Содержимое памяти может быть вызвано даже по частичному входу или искаженному содержанию. Ассоциативная память может найти применение при создании мультимедийных информационных баз данных.

Управление. Примером является оптимальное управление двигателем, рулевое управление на кораблях, самолетах.

3.3.2. Как работает нейронная сеть

Предположим, что нам даются наборы чисел (входные векторы), и для каждого из них нам сообщают значение функции, которое она имеет на данном наборе. Пример: значением является обменный курс некоторой валюты на следующий день, вход – уровень этого курса и некоторых других финансовых показателей за, скажем, последний месяц. Другой пример: входной вектор – характеристики заемщика банка (возраст фирмы, капитал, количество занятых, подвергался ли судимости директор и т. п.), результат – выполнил ли клиент условия возврата кредита. В обоих случаях речь идет пока об исторических данных. Затем нам предъявляют уже новые данные: значения финансовых показателей по сегодняшний день включительно или данные о новом клиенте, обратившемся с просьбой о предоставлении кредита. Резуль-

тат теперь неизвестен, и мы должны его (хотя бы приближенно) найти: каким будет обменный курс завтра, перспективен ли для банка данный клиент.

Как действует в этой ситуации нейронная сеть? Элементарная операция, которую она производит с данными, состоит в следующем: берется «взвешенная» сумма входных величин (т. е. сумма, взятая с некоторыми коэффициентами, которые называются весами). Затем полученная величина преобразуется с помощью нелинейной монотонной функции (функции активации) так, чтобы получившееся в результате значение лежало в интервале от 0 до 1. Описанная конструкция называется искусственным нейроном. Сеть состоит из многих таких нейронов, причем часть из них обрабатывает непосредственно входные данные (первый слой нейронов), другие – сигналы, полученные на выходе с нейронов первого слоя и т. д. (скрытые слои нейронов), и, наконец, есть единственный выходной нейрон, который и выдает нам результат. При этом веса, соответствующие различным нейронам (а иногда и параметры функций активации), могут меняться независимо друг от друга. Обрабатывая исторические (обучающие) данные и меняя при этом веса, сеть стремится наилучшим образом приспособить свой выходной сигнал к известному результату. Этот процесс называется обучением сети. После того как оно закончено, на вход сети можно подать новые данные, и она выдает свой прогноз [9].

3.3.3. Общая схема анализа данных с помощью нейронных сетей

Общая схема анализа данных с помощью нейронных сетей состоит из 5 этапов.

Выбор типологии сети. Существует 9 типов сетей, на этом этапе подбирается наиболее подходящий под задачу тип сети.

Экспериментальный подбор характеристик сети. После выбора типа необходимо подобрать структуру сети (количество нейронов, их веса, взаимосвязи и т. д.).

Экспериментальный подбор параметров обучения. Далее необходимо экспериментально определить параметры обучения: максимальное время обучения, количество данных, максимально допустимую ошибку и т. д.

Обучение сети. По обучающей выборке проводится обучение сети. Предполагается, что обучающая выборка содержит в себе информацию, которая характеризует данные в целом.

Проверка адекватности обучения. Проводится анализ полученных результатов на данных, которые не входили в обучающую выборку. Осуществляется ручной контроль результатов работы нейронной сети.

3.3.4. Пример задач, которые ставятся перед нейронными сетями

В работе А. Горбаня [10] приводится следующий пример использования нейронных сетей: нейронная сеть обучалась предсказывать результаты выборов президента США по ряду экономических и политических показателей. Обученные сети были минимизированы по числу входных параметров и связей. Оказалось, что для надежного предсказания исхода выборов в США достаточно знать ответы всего на пять вопросов, приведенных ниже в порядке значимости:

1. Была ли серьезная конкуренция при выдвижении от правящей партии?

2. Отмечались ли во время правления существенные социальные волнения?
3. Был ли год выборов временем спада или депрессии?
4. Произвел ли правящий президент значительные изменения в политике?
5. Была ли в год выборов активна третья партия?

От использования остальных признаков нейронная сеть отказалась. Более того, эти пять «симптомов» политической ситуации в стране входят в распознающее правило двумя «синдромами». Пусть ответы на вопросы кодируются числами: +1 – «да» и 1 – «нет». Первый синдром есть сумма ответов на вопросы 1, 2, 5. Его естественно назвать синдромом политической нестабильности (конкуренция в своей партии, плюс социальные волнения, плюс дополнительная оппозиция). Чем он больше, тем хуже для правящей партии. Второй синдром – разность ответов на вопросы 4 и 3 (политическое новаторство минус экономическая депрессия). Его наличие означает, что политическое новаторство может, в принципе, уравновесить в глазах избирателей экономический спад. Результаты выборов определяются соотношением двух чисел – значений синдромов. Простая, но достаточно убедительная политологическая теория, чем-то напоминающая концепцию то ли Макиавелли, то ли Ленина («единство партии прежде всего, оно является важнейшим слагаемым политической стабильности») [10].

Этот пример показывает несколько достаточно важных вещей с точки зрения использования нейронных сетей:

1. Нейронные сети могут выделять значимые факторы.
2. Факторы могут быть сгруппированы в «синдром» (см. факторный анализ).
3. Исследование осмысленности работы нейронной сети остается за исследователем.

3.3.5. Минусы подхода

Основным минусом нейронных сетей является то, что процесс обучения нейронной сети и процесс принятия решений абсолютно неконтролируемы. Другими словами, нейронная сеть представляет из себя «черный ящик», на входе которого подаются данные, а на выходе получается результат. Что делает внутри себя нейронная сеть, понять невозможно, поскольку анализируются тестовые данные (происходит обучение нейронной сети), при этом система старается минимизировать ошибку, автоматически изменяя внутренние параметры (веса). Несомненно, получить значения весов в обученной сети возможно, но единственный ответ, который могут дать эти веса, такой: какой параметр в тестовых данных играет какую роль, какую степень важности он имеет. Никаких объяснений относительно смысла этих ролей с помощью этого метода получить невозможно.

3.4. Деревья решений

Первые идеи создания деревьев решений восходят к работам Ховленда (Hoveland) и Ханта (Hunt) конца 50-х годов XX века. Однако основополагающей работой, давшей импульс для развития этого направления, явилась

книга Ханта (Hunt E.B.), Мэрина (Marin J.) и Стоуна (Stone P.J.) «Experiments in Induction», увидевшая свет в 1966 г.

Деревья решений — это способ представления правил в иерархической последовательной логической структуре, который позволяет соотнести объект или ситуацию на входе с одним или несколькими выходными (терминальными) узлами [11]. Под правилом понимается логическая конструкция, представленная в виде «если... то».

Рассмотрим следующую задачу: необходимо построить решающее правило по выдаче кредита физическим лицам. В этом случае дерево решений может выглядеть следующим образом (рис. 3):

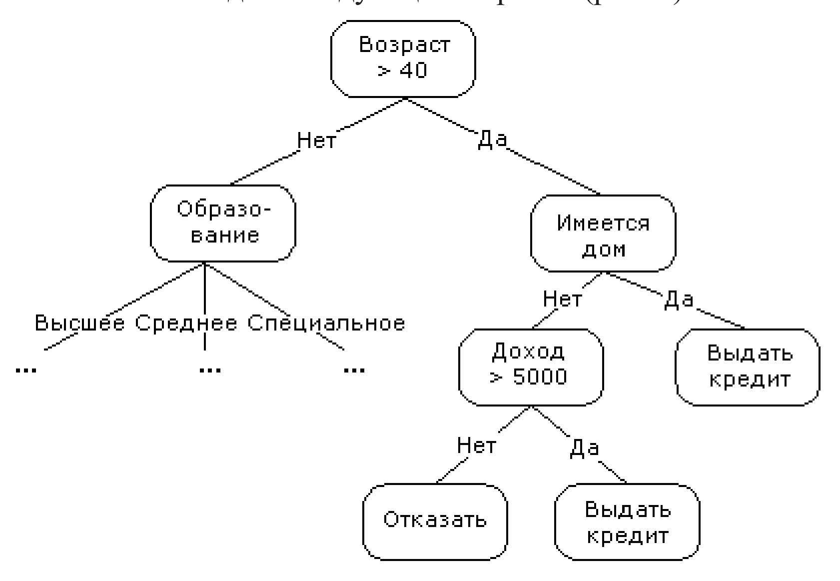


Рис. 3. Четкое дерево решений

В предыдущем примере приведено так называемое *четкое дерево решений*. Неменьший интерес представляют *вероятностные деревья решений* (рис. 4), в которых каждый параметр принятия решения может входить в результирующее решение с некоторой вероятностью [12]:

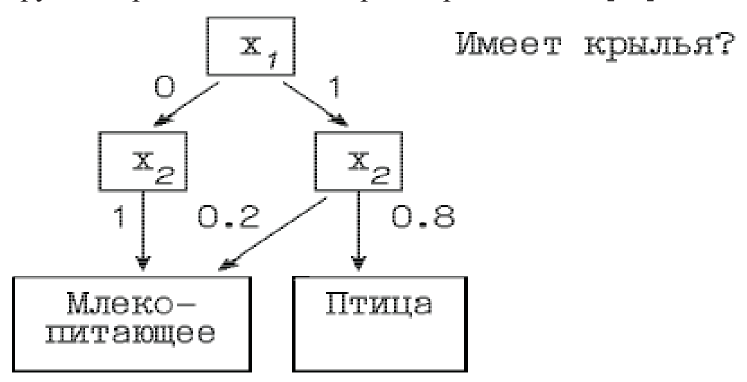


Рис. 4. Вероятностное дерево решений

Метод деревьев решений может помочь при принятии сложного решения, на которое влияют десятки параметров.

Деревья решений широко применяются во многих областях деятельности:

1. *Банковское дело.* Оценка кредитоспособности клиентов банка при выдаче кредитов.

2. *Промышленность.* Контроль за качеством продукции (выявление дефектов), испытания без разрушений (например, проверка качества сварки) и т. д.

3. *Медицина.* Диагностика различных заболеваний.

4. *Молекулярная биология.* Анализ строения аминокислот.

5. *Консалтинг.* Компания McKinsey использует деревья решений (issue tree, термин McKinsey) для консультаций своих клиентов.

Это далеко не полный список областей, где можно использовать деревья решений. Не исследованы еще многие потенциальные области применения этого инструмента.

3.5. Регрессионный анализ

Основной целью регрессионного анализа является определение наличия и характера связи между переменными (в простейшем случае строится зависимость y(x) исходя из примерной формы кривой).

Несколько лет назад американский Институт стратегического планирования провел исследование «Маркетинговая стратегия и уровень прибыли», в котором рассматривалось влияние наиболее значимых переменных на уровень прибыли компании. Выяснилось, что график зависимости рентабельности от доли рынка выглядит следующим образом (рис. 5):

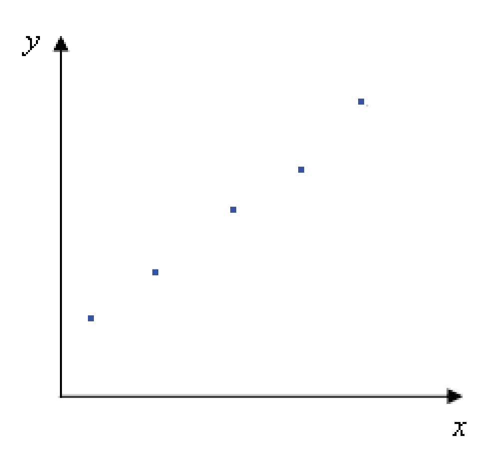


Рис. 5. График зависимости рентабельности от доли рынка

Невооруженным взглядом видно, что это прямая, однако точные ее параметры помогает установить регрессионный анализ.

Регрессионный анализ широко используется в офисном пакете Excel, который предоставляет возможность исследовать не только линейные, но и

другие, более сложные зависимости (в Excel это называется построением линий трендов).

3.6. Дискриминантный анализ

Дискриминантный анализ – это инструмент статистики, который используется для принятия решения о том, какие переменные разделяют возникающие наборы данных.

Например, некий исследователь в области образования решает исследовать, какие переменные относят выпускника средней школы к одной из трех категорий: (1) поступающий в колледж, (2) поступающий в профессиональную школу или (3) отказывающийся от дальнейшего образования или профессиональной подготовки. Для этой цели исследователь может собрать данные о различных переменных, связанных с учащимися школы: пол, возраст, успеваемость, материальное положение семьи и т. д. После выпуска большинство учащихся должно попасть в одну из названных категорий.

Затем можно использовать дискриминантный анализ для определения того, какие переменные дают наилучшее предсказание выбора учащимися дальнейшего пути. Например, можно математически определить, что учащиеся с низкой успеваемостью и низким достатком в семье скорее всех попадают в 3-ю категорию.

Еще пример [13]: есть данные о клиентах / потребителях, которых можно разделить по группам (совершившие повторную покупку – не совершившие повторную покупку; покупатели марки А – покупатели марки В – покупатели марки С; высокие риски невозврата кредита – низкие риски невозврата кредита), также имеется дополнительная информация о клиентах / потребителях. Дискриминантный анализ позволяет выяснить, действительно ли группы различаются между собой, и если да, то каким образом (какие переменные вносят наибольший вклад в имеющиеся различия).

3.7. Корреляционный анализ

Корреляционный анализ позволяет судить о том, насколько похоже ведут себя разные переменные [14]. В самом общем виде принятие гипотезы о наличии корреляции означает, что изменение значения переменной А произойдет одновременно с пропорциональным изменением значения Б: если обе переменные растут, то корреляция положительная; если одна переменная растет, а вторая уменьшается – корреляция отрицательная.

При изучении корреляций стараются установить, существует ли какая-то связь между двумя показателями в одной выборке (например, между ростом и весом детей или между уровнем IQ и школьной успеваемостью) либо между двумя различными выборками (например, при сравнении пар близнецов), и если эта связь существует, то сопровождается ли увеличение одного показателя возрастанием (положительная корреляция) или уменьшением (отрицательная корреляция) другого.

Заключение

Задачи восстановления зависимостей активно изучаются уже более 200 лет, с момента разработки К. Гауссом в 1794 г. метода наименьших квадратов. В математической статистике с этого времени было разработано огромное количество методов и инструментов анализа данных. В данной работе описаны методы, которые наиболее широко используются во всем мире и во всех областях прикладной науки для анализа данных – физике, биологии, экономике, психологии и др.

В настоящее время компьютеры играют большую роль в математической статистике. Они используются как для расчетов, так и для имитационного моделирования (в частности в методах размножения выборок и при изучении пригодности асимптотических результатов. Дополнительную информацию по методам анализа данных и математической статистике можно получить в работах [15;16].

Литература

1. Ефремова М.В. Сегментация потребителей гостиничных услуг // Маркетинг в России и за рубежом. 2002. № 2.
2. Толстова Ю.Н. Измерение в социологии: Курс лекций. М.: ИНФРА-М, 1998.
3. Гаскаров Д.В., Шаповалов В.И. Малая выборка. М.: УРСС, 1978 // [Электронный ресурс] http://ru.wikipedia.org/wiki/Выборка
4. Tryon R.C. Cluster Analysis. NY.: McGraw-Hill, 1939.
5. Дюран Б., Оделл П. Кластерный анализ / Пер. с англ. М.: УРСС, 1977.
6. Садохина Е.Ю. Факторный анализ структуры экономического роста России и Беларуси за 1991–2002 гг. // Научные труды ИНП РАН.
7. McCulloch W.S. and Pitts W. A logical Calculus of Ideas Immanent in Nervous Activity // Bull. Mathematical Biophysics. 1943. Vol. 5. P. 115–133.
8. Сергей Колесников. Зачем нужны нейронные сети? // [Электронный ресурс]КОМПЬЮТЕР–ИНФОРМ, http://www.ci.ru/inform15_05/p_08.htm
9. Курочкин С.В. Нейронные сети: просто о сложном. Теория вероятностей и ее применения // [Электронный ресурс] http://www.tvp.ru/prog/kur_neur.htm
10. Горбань А. Нейроинформатика и ее приложения // [Электронный ресурс] gorban@cc.krascience.rssi.ru (Вычислительный центр СО РАН, Красноярск-36.)
11. Акобир Шахиди. Деревья решений – общие принципы работы // [Электронный ресурс] http://www.basegroup.ru/trees/
12. http://www.dmitry-kazakov.de/post_gra/chapter2.htm
13. Позднякова А. Дискриминантный анализ и другие многомерные методы в маркетинговых исследованиях / CIU (Consumer Insights Ukraine). Киев.

14. [Электронный ресурс] http://ru.wikipedia.org/wiki/Корреляционный анализ
15. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. М.: Наука, 1983.
16. Вероятность и математическая статистика: Энциклопедия / Гл. ред. Ю.В. Прохоров. М.: Изд-во «Большая Российская Энциклопедия», 1999.